# CIVIC DIGITAL FELLOWSHIP

# Linking Export Declarations to Commodity Flow Survey Data

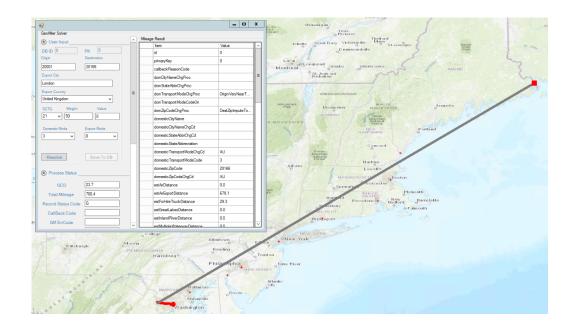
By: Anna Kong Economic Reimbursable Surveys Division Supervisor: Emily Ross





### **Background**

- The Commodity Flow Survey is sent to over 100,000 businesses that ship commodities.
- GeoMiler takes in survey data and outputs important attributes such as port-of-exit used for mileage calculations of freight.
- With the Census moving away from GeoMiler, alternative methods of calculating port-of-exit are needed.
- The aim of this project is to link CFS export shipments to their associated Foreign Trade Division (FTD) export declaration.
- To do this, we built a machine learning model to match records between the two datasets.



# **Project Benefits**

#### **Higher data quality**

- Eliminate redundant questions.
- Save time for respondents.
- Linking datasets improves accuracy, helps validation.

#### **Saving costs**

- The cost to maintain GeoMiler is high.
- A machine learning system only takes a few hours to run and uses existing infrastructure and computing power.
- We calculate savings of at least \$600k/yr.
   for ERD and this is a low estimate.





## **Project Status**

Complete

- Data cleaning and harmonization of ~5.7 million rows
- Hand labeled records
- Generating vectors of similarity scores
- Classification framework and training initial model.

Inprogress

- Testing and improving accuracy
- Scaling up classification to full set of records
- Validation and evaluation of results.

Next Steps

- Moving results into imputation
- Active learning pipeline for labeling matches
- Comparison with alternative methods and/or packages

Shape your future START HERE >



# **Looking Forward**

 The work done with matching records is proof of concept of a bigger picture idea: how can we combine data across silos to take advantage of information we have bureau wide?





# Thank you!



